MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 2, 2018/2019

TNL3221 – NATURAL LANGUAGE PROCESSING

(All Sections / Groups)

12 MARCH 2019
2.30pm – 4.30pm
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This question paper consists of 5 pages with 4 questions only.
2. Attempt **ALL** questions. All questions carry equal marks and the distribution of the marks for each question is given.
3. Please write all your answers in the Answer Booklet provided.

QUESTION 1

(a) Differentiate between stemming and lemmatization. Assuming that stemming rule "ies → i" is used, show the output for the word "studies" after stemming and lemmatization process, respectively. [3 marks]

(b) Briefly describe the three phases in information retrieval (IR)-based factoid question answering. [3 marks]

(b) Suppose a corpus contains 400000 word tokens and 80000 of these are tagged as N (common noun). The word form "cook" occurs 1000 times in the corpus, tagged either as N or V. Analysis shows that "cook" accounts for 0.4% of all common noun tokens in the corpus. Use Bayes' formula to calculate the probability that a given occurrence of "cook" is tagged as N. [4 marks]

QUESTION 2

(a) Write a regular expression for each of the following: [6 marks]

i. That matches the following strings: Artificial Intelligence, Artificial intelligence, artificial Intelligence, artificial intelligence.
ii. That matches the following strings: AI, ai, A.I.
iii. That matches prices (examples: RM199, RM123.88).
iv. That matches an arithmetic expression using integers, addition and multiplication (examples: 2+3, 5+9*1, 47*6+2, 88*9+73*56).
v. That matches clitic words (examples: doesn't, we're)
vi. That matches hyphenated compound words (examples: 12-point, well-being, merry-go-round, state-of-the-art).

(b) Given the contingency table below: [4 marks]

i. Calculate Precision.
ii. Calculate Recall.
iii. Calculate Accuracy.
iv. Calculate F-measure with β=2.

Show your final results to three decimal places.

		Gold Labels	
		True	False
System Output	True	278	345
	False	89	47

Continued...

QUESTION 3

(a) Consider the following grammar:

S → NP VP S → S Conj S NP → Det N NP → NP Conj NP VP → V VP → V NP VP → V NP PP PP → P NP	Det → the | a | some N → police | constable | burglar | burglars | bed V → called | gave | has | have | arrest | arrests | is | are P → in | on | under | by Conj → and | or

i. Draw a tree structure for the sentence "the police and the constable arrest a burglar under the bed", according to the grammar rules above. [2 marks]

ii. Show the shift-reduce parsing of the sentence "the police and the constable arrest a burglar under the bed". [3 marks]

(b) Based on the Levenshtein distance with insertion cost 1, deletion cost 1 and substitution cost 2:

i. Compute the edit distance of "relavant" to "elephant". Show your work using the edit distance grid. [2 marks]

ii. Compute the edit distance of "relavant" to "relaxant". Show your work using the edit distance grid. [2 marks]

ii. As a spelling checker, suggest a closest word for "relavant" based on the computed edit distance. [1 mark]

Continued...

QUESTION 4

(a) Given the tag transition probabilities in Table 1, word likelihood probabilities in Table 2 and the part-of-speech tags: Janet/NNP, back/VB, the/DT and bill/NN.

i. List the possible part-of-speech tags for the word "will". [1.5 marks]

ii. Calculate and justify the best tag for the word "will". [2.5 marks]

	NNP	MD	VB	JJ	NN	RB	DT
<s>	0.2767	0.0006	0.0031	0.0453	0.0449	0.0510	0.2026
NNP	0.3777	0.0110	0.0009	0.0084	0.0584	0.0090	0.0025
MD	0.0008	0.0002	0.7968	0.0005	0.0008	0.1698	0.0041
VB	0.0322	0.0005	0.0050	0.0837	0.0615	0.0514	0.2231
JJ	0.0366	0.0004	0.0001	0.0733	0.4509	0.0036	0.0036
NN	0.0096	0.0176	0.0014	0.0086	0.1216	0.0177	0.0068
RB	0.0068	0.0102	0.1011	0.1012	0.0120	0.0728	0.0479
DT	0.1147	0.0021	0.0002	0.2157	0.4744	0.0102	0.0017

Table 1: Tag transition probabilities. Rows are labeled with the conditioning event; thus P(VB|MD) is 0.7968.

	Janet	will	back	the	bill
NNP	0.000032	0	0	0.000048	0
MD	0	0.308431	0	0	0
VB	0	0.000028	0.000672	0	0.000028
JJ	0	0	0.000340	0.000097	0
NN	0	0.000200	0.000223	0.000006	0.002337
RB	0	0	0.010446	0	0
DT	0	0	0	0.506099	0

Table 2: Word likelihood probabilities.

(b) Given the following short product reviews as the training set, each labelled with a class, either positive or negative:

Review	Text	Class
1	excellent quality low cost	+
2	good quality fast delivery	+
3	good packaging low price	+
4	acceptable quality low price	+
5	bad product lousy quality	-
6	low quality slow delivery	-

i. Assume a naive Bayes classifier and use add-1 smoothing for the likelihoods. Compute the most likely class for the test review D "low bad quality product" [3 marks]

ii. Justify the class of the test review D. [1 mark]

Continued...

(c) Given the following three sample queries for an information retrieval system. In each case, the system returns search results ranked in the confidence level.

i. Calculate the mean reciprocal rank of the system. [1 mark]

ii. Assumed that the information retrieval system receives the fourth query and returns no result for the query, what is the mean reciprocal rank of the system? [1 mark]

Query	**Results**	**Correct response**
Finance accounting	1. Financial statement 2. Business accounting 3. Financial accounting 4. Warren Buffett accounting	Financial accounting
Language processing	1. Language processing disorder 2. Natural language processing 3. Speech and language processing 4. Language processing in the brain	Natural language processing
MMU	1. Multimedia University 2. Manned Maneuvering Unit 3. Manchester Metropolitan University 4. Multi Material Upgrade	Multimedia University

End of Paper